M-Sieve: A visualisation tool for supporting network security analysts

VAST 2012 Mini Challenge 1 Award: "Subject Matter Expert's Award"

S. Choudhury, N. Kodagoda, P. Nguyen, C. Rooney, S. Attfield, K. Xu, Y. Zheng, B.L.W. Wong, R. Chen, G. Mapp, L. Slabbert, M. Aiash, A. Lasebae

Middlesex University

ABSTRACT

The Middlesex Spatial Interactive Visualisation Environment (M-Sieve) is a spatiotemporal visual analytics tool for exploring computer network activity. M-Sieve allows the user to filter and visualize data through facets to explore and find patterns. To help guide exploration, we developed a set of rules which are used to derive a variable we call the 'Concern Level Assessment' (CLA). The CLA is based on attributes of nodes on the network. The rules were developed by eliciting inferences from network security domain experts. The combination of M-Sieve and the CLA allowed us to address the problem presented by the VAST 2012 Competition - Mini Challenge 1.

Keywords: Visual Analytics, Information Visualisation, Network Security, Human Factors.

Index Terms: H.5.2 [User Interfaces]: Information Visualization; C.2.3 [Network Operations]: Network Management

1 INTRODUCTION

The IEEE VAST 2012 Mini-Challenge 1 (MC 1) [1] set the task of investigating activity within the computer network of the Bank of Money (BOM), which stretches over the whole of BankWorld (including multiple time zones). The BOM network has 888,977 machines. The data provided for the problem comprises metadata indicating the health of each machine sampled at fifteen-minute intervals over a 48-hour period (192 time points in all).

In this paper we describe our solution to the challenge. We begin by describing the use of off-the-shelf tools for initially exploring and storing the raw data. We then describe the M-Sieve tool and the Concern Level Assessment (CLA). Finally, we give examples of how these were used to identify problems in BOM's network.

2 OFF-THE-SHELF TOOLS

The off-the-shelf tools we used were:

- *MS Excel* to view and drill down to the raw data.
- *SPSS* for deriving central tendency and variability statistics for numbers of connections in given contexts (machine class, function and time of day). This used in the derivation of the CLA.
- *MySQL* for storing the raw data.

{t.choudhury}{n.kodagoda}{p.nguyen}{c.rooney}
{s.attfield}{k.xu}{y.zheng}{w.wong}{r.chen}{g.mapp}
{l.slabbert}{m.aiash}{a.lasebase} @mdx.ac.uk

3 CONCERN LEVEL ASSESSMENT (CLA) RULES

To better interpret the patterns within the data, and distinguish between the important and anomalous we developed a machine inferred variable called 'Concern Level Assessment' (CLA) which is based on parameters within the data. To establish a basis for calculating CLA we first conducted a series of knowledge elicitation interviews with network security experts using samples from the data as context. We then used findings from the interviews to develop a set of 97 inference rules that could be used to perform a cursory, automated interpretation of the data. The rules embody abductive inferences from parameters including machine type (class and function), policy status, activity flag, number of connections (statistically determined to be either normal or abnormal based on a given class and function of machine) and time of day (during or after business hours).

We used the rules to derive a CLA on a six-point scale: *normal, low, medium-low, medium, medium-high, and high*. *Normal* indicates that a machine is of no concern (e.g., the machine has activity flag 1, policy status 1 and has a number of connection that is within the mean + 1 standard deviation for that machine class for that time of day), whilst *high* indicates a machine is of significant concern (e.g., a machine that has a virus or a machine with an abnormally high level of connections for that machine type and is suffering from 100% CPU consumption and it is afterhours). In addition, we derive a CLA of *very high* where a machine has policy status five and activity flag five since this indicates that the machine is both infected with a virus and has an external device added (this is of extra concern as it may result in a rapid spread of infection). We aim to publish furthers details the CLA, including the rules and how they were generated, in the near future.

4 M-SIEVE

In this section we describe the Middlesex Spatial Interactive Visualisation Environment. We use the name M-Sieve to denote an ease for quickly and easily filtering (or sieve) large amounts of data.

The M-Sieve interface Figure 1 has three areas. To the left is a map of BankWorld (a) with overlays indicating the different locations of machines. Colour is used to indicate the maximum policy status of machines at each location, shape indicates location type (branch, data centre, etc.), and size indicates the number of machines at a location (on a user definable scale). The regions are overlaid on top, and a white or black background symbolises whether or not the time at that location is within business hours.

The right-hand side of the interface contains an attribute explorer (b), which represents each attribute as an interactive histogram [2]. The attribute explorer shows the distribution of machines over a vertically arranged set of horizontal histograms, where each histogram corresponds to an attribute extracted from the dataset (from top to bottom): *policy status*, *activity flag*,

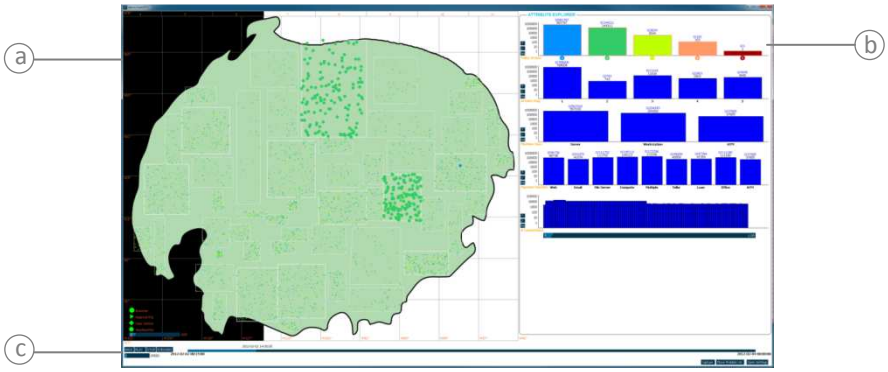

Figure 1: The M-SIEVE application; a) the spatial window, b) the attribute explorer, c) timeline control.

machine class, *machine function* and *number of connections*. These are visualised on a logarithmic scale.

Selecting a column on a histogram creates a filter corresponding to the associated bin range. Multiple column selections on a given histogram create an OR query using the selected ranges. Selecting bins across histograms creates an AND query between histograms. On selection, the query is automatically executed and the histograms are dynamically updated to show the distribution of values on the subset.

Using multiple coordinated views [3] updates in the attribute explorer are automatically reflected in the map view. Conversely, regions can be selected within the map to further subset the data.

The final part of the interface is a time-bar (c). Clicking on the bar moves an indicator to that time-point and loads the corresponding status of machines according to the current filter. A play/pause button can also be used to automatically play through the dataset, with the map view and histograms synchronously updating with each time point.

The UI also contains further menu options, including the ability to view the CLA of each individual machine within the current filter (i.e. drill-down). This is presented as a table, but can also be export to other visualisations. For example, Figure 2 shows a TreeMap [4] visualisation of the concern level across all machines at 2pm BMT on the 2nd of February.

5 Example Of Use

Using M-Sieve and the CLA, we identified a number of anomalous events within the dataset. In a real life scenario, a real-time system would provide a network security analyst with a CLA alert of descending priority. For example, at time point: 2pm BMT, 2nd February, the following alerts would be provided:

- **1 machine at CLA High** – Machine requires immediate action. For example, server 172.2.194.20 in HQ datacenter-2, Region 36 is rated high due to virus infection.
- **811 machines at CLA Medium-High** – Machines require attention, but the reason is not conclusive. For example, Web Server 172.8.28.77 has activity flag 4 and a statistically high number of connections for a web server. This can indicate a denial of service attack.
- **12,065 machines at CLA Medium** – Machines need to be monitored to see whether they escalate over time. For example, some servers in the Region 10 Headquarters are rated CLA medium due to the number of login failures. These are expected for workstations but not for servers.

Another example of how the CLA helped us was in the investigation of machines at the headquarters in Region 10, where all have at least some level of concern. The majority are rated medium-low, but we determine that the facility as a whole is a cause for concern. The majority of the low and medium-low concern levels were due to machines with a policy status other than 1. These concerns are also flagged for a high number of connections, login failures, and high CPU consumption on specific machines types.

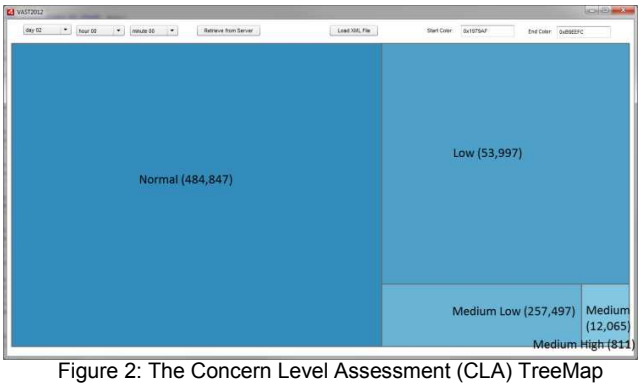

Figure 2: The Concern Level Assessment (CLA) TreeMap

6 Conclusion & Future Work

M-Sieve, a tool for performing spatiotemporal network analysis, was developed to visualise and investigate the VAST 2012 Mini Challenge one dataset. In the process, interesting patterns were discovered that were then interpreted using CLA rules generated with the help of domain experts. We hope to continue to improve both M-Sieve and the CLA, and further investigate how they improve the task of network security analysis. We would also like to evaluate how well our tool can improve situational awareness when applied to domains other than network security.

7 References

[1] K. Cook, G. Grinstein, and M. Whiting, "VAST Challenge 2012: Challenge Descriptions," *VAST2012*, 2012. [Online]. Available: http://www.vacommunity.org/tiki-index.php?page=VAST Challenge 2012: Challenge Descriptions. [Accessed: 13-Aug-2012].

[2] L. Tweedie, B. Spence, D. Williams, and R. Bhogal, "The Attribute Explorer," in *CHI '94*, 1994, pp. 435–436.

[3] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky, "Guidelines for using multiple views in information visualization," *Proceedings of the working conference on Advanced visual interfaces - AVI '00*, pp. 110–119, 2000.

[4] B. Johnson and B. Shneiderman, "Tree-maps: a space-filling approach to the visualization of hierarchical information structures," *Proceeding Visualization '91*, pp. 284–291, 1991.